A határozott és határozatlan ragozás hibáinak automatikus felismerése magyarul tanulók szövegeiben

Vincze Veronika[1], Zsibrita János[2], Durst Péter[3], Szabó Martina Katalin[4]

[1] MTA-SZTE Mesterséges Intelligencia Kutatócsoport
vinczev@inf.u-szeged.hu

[2] Szegedi Tudományegyetem, Informatikai Tanszékcsoport
zsibrita@inf.u-szeged.hu

[3] Szegedi Tudományegyetem, Hungarológia Központ
durst.peter@gmail.com

[4] Szegedi Tudományegyetem, Magyar Nyelvészeti Tanszék
szabomartinakatalin@gmail.com

Kivonat: Jelen munka célja, hogy a HunLearner magyar nyelvtanulói korpuszban automatikusan azonosítsuk a határozott és határozatlan igeragozásban elkövetett nyelvtanulói hibákat. A hibaelemzés rámutat a nyelvtanulók számára nehézséget okozó nyelvtani szerkezetekre, ami az adott jelenségek célzott oktatásában és gyakorlásában hasznosítható a nyelvoktatás felől nézve, számítógépes oldalról pedig egy nyelvhelyesség-ellenőrző továbbfejlesztésében lehet hasznos.

1 Bevezetés

A jelen dolgozatban a HunLearner magyar nyelvtanulói korpuszban [1] folyó munkálatok egyik részfeladatáról számolunk be. A projekt a határozott és határozatlan igeragozásban elkövetett nyelvtanulói hibák automatikus azonosítását tűzte ki célul. Az általunk vizsgált ragozásnak több elnevezése is elterjedt (*tárgyas ragozás*, *határozott ragozás*, *határozott tárgyas ragozás*, vö. [2]), ebben a dolgozatban *a határozott tárgyas ragozás* terminust használjuk.

Munkánkban először röviden ismertetjük a határozott és határozatlan tárgyak típusait. Ezek után bemutatjuk a vizsgálatunk alapjául szolgáló HunLearner korpusz bővített változatát, majd megmutatjuk, miként lehetséges automatikus eszközökkel azonosítani a határozott ragozásban elkövetett hibákat. A leggyakoribb hibatípusokról végül statisztikai elemzéseket is adunk.

2 Határozott tárgyak

A magyar nyelv sajátosságai közül kiemelkedik a határozott tárgyas ragozás, amely kifejezetten kevés nyelvben figyelhető meg. Széles körben elterjedt elnevezése a rövidebb tárgyas ragozás terminus, a grammatikák azonban inkább határozott tárgyas ragozásként említik [2]. A határozott igei paradigma használati szabályainak elsajátítása és alkalmazása gyakran okoz nehézséget a magyar nyelv tanulói számára, ráadásul a határozott tárgy különböző típusai is eltérő mértékben okoznak nehézséget a nyelvtanulás során. A határozott ragozást a struktúrában megjelenő ún. határozott tárgy hívja elő, tehát a tárgy határozottságát jelölni kell az igén. Ezt harmadik személyű tárgyakkal tudjuk kifejezni teljes paradigmában, a második személyű tárgyak jelölésére csak hiányos ragozási sor áll rendelkezésre a magyarban (vö. *ismerem őt*, *ismered őt* vs. *ismerlek téged*).

A határozott ragozás több nyelvi szinten átívelő jelenség, amelynek lényegét M. Korchmáros nyelvtanában [3] így foglalja össze: „Általában akkor beszélünk a magyar igeragozás szempontjából megkülönböztetett határozott tárgyról, ha az a beszélő és a hallgató tudatában egyforma mértékben azonosított egyedi vagy annak tekintett objektum(ok)at jelöl.” Ez az egyébként nagyon pontos megfogalmazás azonban még nem ad elég fogódzót sem a magyar nyelv határozott tárgyas ragozásának elsajátításához, sem pedig annak számítógépes feldolgozásához; mindenképpen szükség van a határozott tárgyas ragozást megkövetelő határozott tárgyi tömbök pontos és részletes bemutatására. A leggyakoribb és a nyelvtanulók számára is a legkisebb nehézséget jelentő határozott tárgyak a következők:

1. A tárgy tulajdonnév:
Ismerem Klárit.

2. A tárgy határozott névelővel álló névszó:
Megesszük az almát.
Elviszem a pirosat.

3. A tárgy főnévi mutató névmás:
Ezt kérem.

4. A tárgy birtokos személyjellel vagy *-é* birtokjellel álló névszó:
Mindenki ismeri a testvéremet.
A Katiét vették meg.

5. A tárgy visszaható / kölcsönös / birtokos névmás:
Mindenki magát látja a tükörben.
Szeretik egymást.
A mienket ne vidd el.

6. A tárgy harmadik személyű személyes névmás:
Ismerem őt.

Érdekesség, hogy a személyes névmások közül csupán a harmadik személyűek számítanak határozott tárgynak, hiszen a határozott tárgyas ragozás alapvetően csak harmadik személyű tárgyra tud utalni.

7. A tárgy *-ik* kijelölő jellel áll:
Csak az egyiket kérem.
Melyik könyvet olvastad?
Hányadikat eszed már?

Meg kell jegyezni, hogy a *Melyik?* és a *Hányadik?* kérdőszón kívül más kérdő névmás nem minősül határozott tárgynak.

8. A tárgy egy mellékmondat:
Tudom (azt), ki vagy.

A tárgyi alárendelő mellékmondatok több formában is előfordulhatnak, hiszen a főmondatban nem jelenik meg szükségszerűen az *azt* utalószó. Ez a változatosság mind a nyelvtanulók, mind a számítógépes nyelvfeldolgozás szempontjából igen problematikusnak tekinthető.

9. A tárgy a *mind* vagy a *valamennyi* névmás:
Mind elolvasta.
Valamennyit megették.

A *valamennyi* névmást illetően fontos hangsúlyozni, hogy az csupán annak 'összeset' jelentésében jár határozott ragozással. Ennek következtében a szerkezet használatának elsajátítását tovább nehezíti, hogy esetében csak a szövegkörnyezet segítségével lehet eldönteni, hogy milyen ragozást kell használni.

10. **A tárgy explicit módon nem jelenik meg a mondatban:**
Add ide!
Tegnap vettünk egy esernyőt. Ma elvesztettük.

Az explicit módon nem realizálódó határozott tárgy főleg a párbeszédes formájú szövegekre jellemző, és, mivel az adott szerkezetben fonológiailag nem realizálódik, az adott kontextus mutatja meg a szerkezetben való létezését. Ilyenkor vagy egy a szövegben már korábban említett, vagy pedig egy nyelven kívüli eszközökkel (pl. rámutatás) azonosított tárgyról van szó.

3 Kapcsolódó irodalom

A számítógépes nyelvfeldolgozás szempontjából a határozott tárgy kezelése problematikusnak tekinthető, ugyanis mint láttuk, a határozott tárgyi tömbök morfológiai megjelenése nem egységes, emiatt automatikus felismerésük bizonyos esetekben akadályokba ütközik. A témához kapcsolódó korábbi korpuszalapú kutatások között találunk kínai anyanyelvűekkel végzett, szóbeli mintavételen alapulót [4], eltérő anyanyelvű válaszadókkal végzett kérdőíves tesztelést [5], valamint egy ugyancsak kérdőíven alapuló vizsgálatot homogén, mordvin anyanyelvű csoporttal [6]. Ugyanakkor meg kell említenünk, hogy – a jelen projekttől eltérően – egyik esetben sem használtak még automatikus eszközöket a határozott tárgy, valamint a határozott ragozásban vétett nyelvtanulói hibák feldolgozásának céljából.

4 A HunLearner korpusz

A HunLearner korpusz magyar mint idegen nyelv szakos egyetemi hallgatók fogalmazásait tartalmazza [1]. Horvát anyanyelvű diákok három nagyobb témában írtak esszét: *Egy szimpatikus ember*, *Nehézségek a magyar nyelv tanulásában*, *illetve Magyar bevándorlók Angliában*. A korpuszban a főneveket érintő morfológiai hibákat kézzel javítottuk, és minden hibához automatikusan hozzárendeltük annak típusát.

A korpusz néhány új szöveggel bővült a közelmúltban. Ezeket észt diákok írták az *Egy szimpatikus ember* témában. A korpusz jelen, kibővített változatában 1427 mondat és 22 000 token szerepel.

5 Határozott ragozási hibák a korpuszban

A HunLearner korpusz szövegeit a magyarlanc szoftverrel [7] automatikusan elemeztük, majd a morfológiai és szintaktikai elemzés alapján szabályokat definiáltunk az tárgy-ige egyeztetés különböző típusaira. Ezek alapján automatikusan össze tudtuk gyűjteni azokat az eseteket, amelyekben eltérés mutatkozott a tárgy típusa által indikált és a tényleges igeragozás között. Például: megvizsgáltuk, hogy a köznévi tárgy rendelkezik-e névelővel. Amennyiben rendelkezik határozott névelővel, az igeragozásnak határozottnak kell lennie.

Az alábbi példában a főnévi igenév mutató névmási tárgya határozott ragozást váltana ki a *szeret* igén, azonban a nyelvtanuló határozatlan ragozást használ: *Végül mindenkinek **szeretnék** azt mondani, hogy Angliában tök jobb életem van, mint Magyarországban.*

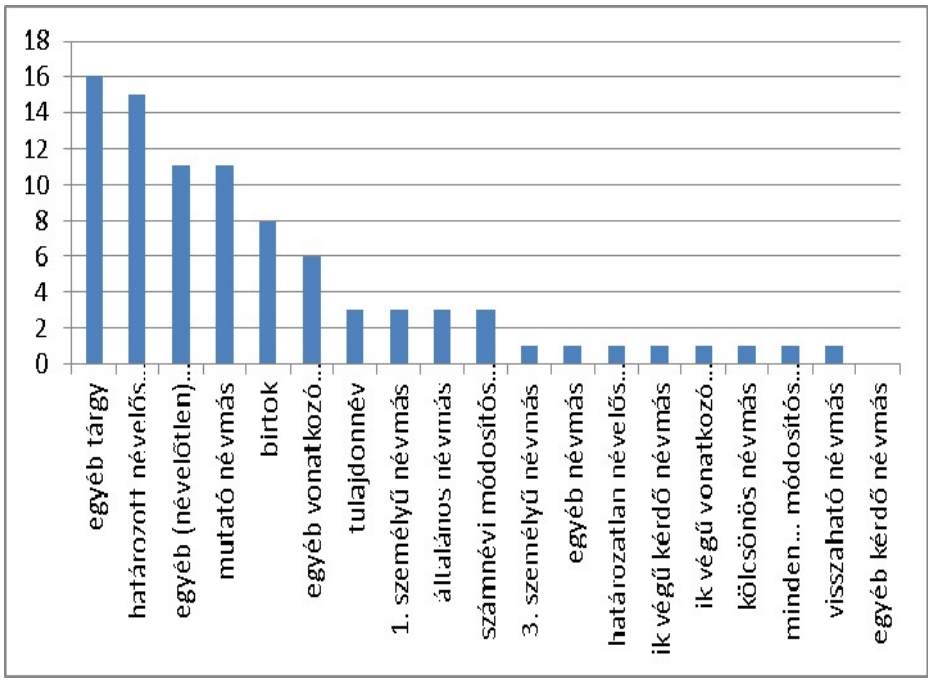
Az 1. táblázat mutatja a vizsgálat számszerű eredményeit. Jelen cikk keretei között csak azokat az eseteket vizsgáltuk részletesebben, ahol a tárgy fonológiailag is jelen van a mondatban (*Van tárgy a mondatban* oszlop), tehát egyelőre nem foglalkozunk azokkal az esetekkel, amikor a névmási tárgy jelenléte pusztán a határozott ragozású igéből lenne kikövetkeztethető. Az alárendelő mellékmondati tárgyakat is kizártuk a vizsgálatból, hiszen a tárgyi szerepet betöltő mellékmondatok automatikus azonosítására jelenleg nem képes a magyarlanc szintaktikai modulja. Kizártuk a vizsgálatból továbbá azokat a morfológiailag többértelmű igealakokat is, ahol a határozott és határozatlan ragozás egybeesik (pl. múlt idő E/1. alakban, vö. *olvastam*), itt ugyanis nem eldönthető, hogy a nyelvtanuló határozott vagy határozatlan ragozást kívánt-e használni.

A szűrések után kapott 87 esetet további vizsgálatoknak vetettük alá. Az eredmények szerint a leggyakoribb hibaforrás a határozott névelős köznévi tárgy: ez határozott ragozást váltana ki, azonban a hibák 17%-ában határozatlan ragozású igével szerepel együtt. Két másik gyakori hiba a mutató névmási tárgy és a névelőtlen köznévi tárgy, melyek a hibák 13-13%-ában a nem megfelelő ragozású igével fordulnak elő. A birtokos személyjellel ellátott tárgyakat érintő hibákat is ideszámítva elmondhatjuk, hogy a határozott ragozást érintő hibák 50%-áért a fenti hibák felelnek.

1. táblázat: Ragozásbeli eltérések.

Alkorpusz	Igék száma	Ragozásbeli eltérés	Van tárgy a mondatban	Egyértelmű igealak
Nehézségek	1018	149	42	32
Anglia	564	74	46	16
Szimpatikus	841	149	47	39
Összesen	2423	372	117	87

Az 1. ábra mutatja a hibásan használt igeragozást kiváltó tárgytípusok gyakoriságát.



1. ábra: Hibás igeragozást kiváltó tárgyak.

Az eredmények egyben azt is mutatják, hogy jóval több a határozott tárgy-határozatlan igealak típusú tévesztés (59%), mint a határozatlan tárgy-határozott igealak típusú.

6 Az eredmények felhasználása

A vizsgálat eredményeit egyrészt kitűnően hasznosíthatja a nyelvoktatás, hiszen a hibák statisztikai elemzése lehetőséget nyújt arra, hogy a nehezebbnek bizonyuló szerkezeteket célzottan gyakorolhassák a diákok a nyelvórán. Másrészt számítógépes

nyelvészeti oldalról nézve az egyeztetési hibák automatikus hibajavítása előtt is megnyílik a lehetőség, hiszen a tárgy típusa alapján meg lehet határozni az elvárt igealakot, és amennyiben nem a megfelelő szerepel a szövegben, egy nyelvhelyesség-ellenőrző program javítási javaslatokat tehet az igealakra nézve.

7 Összegzés

Ebben a munkában bemutattuk számítógépes nyelvészeti eszközökön alapuló megközelítésünket, mely a határozott és határozatlan ragozásban elkövetett hibák automatikus azonosítását célozza. A vizsgálatból kiderült, hogy melyek azok a nyelvtani szerkezetek, amelyek problémát jelentenek a magyart mint idegen nyelvet tanulók számára. Ezen eredmények haszna elsődlegesen a nyelvoktatásban mutatkozik meg, hiszen a nyelvtanulók így célzottan gyakorolhatják a problémásabb szerkezeteket, mindemellett a határozott és határozatlan ragozás hibáinak automatikus azonosítása egy nyelvhelyesség-ellenőrző programban is jó szolgálatot tehet.

Köszönetnyilvánítás

A jelen kutatás a futurICT.hu nevű, TÁMOP-4.2.2.C-11/1/KONV-2012-0013 azonosítószámú projekt keretében az Európai Unió támogatásával és az Európai Szociális Alap társfinanszírozásával valósult meg.

Hivatkozások

1. Vincze V., Zsibrita J., Durst P., Szabó M. K.: HunLearner: a magyar nyelv nyelvtanulói korpusza. In: Tanács A., Vincze V. (szerk.): IX. Magyar Számítógépes Nyelvészeti Konferencia. Szegedi Tudományegyetem, Szeged (2013) 97–105
2. Pete I.: A határozott tárgyas ragozásról. Magyar Nyelvőr, Vol. 130. (2006) 317–324
3. M. Korchmáros V: Lépésenként magyarul. Magyar nyelvtan – Nem csak magyaroknak.. Szegedi Tudományegyetem, Szeged (2006)
4. Langman, J., Bayley, R.: The acquisition of verbal morphology by Chinese learners of Hungarian. Language variation and Change, Vol. 14 (2002) 55–77
5. Durst P.: A magyar főnévi szótövek és egyes toldalékok elsajátításának vizsgálata magyarul tanuló külföldieknél. Hungarológiai Évkönyv, Vol. 11. Pécs (2010)
6. Durst, P., Janurik, B.: The Acquisition of the Hungarian definite conjugation by learners of different first languages. Lähivõrdlusi. Lähivertailuja 21. Tallinn: Estonian Association for Applied Linguistics (EAAL) (2011) 19-44
7. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In: Proceedings of RANLP-2013, Hissar, Bulgaria (2013) 763–771